



Généralités

[Tout replier](#)



[Annonces](#) ○



First part: linear models, evaluation

Topics were:

- Linear models
- Cross-validation
- One-Hot Encoding and dummies
- Interaction effect



Modèles mixtes,...



[Introduction au modèle mixte.pdf](#) PDF

55.6 Ko



[intro_mixte_2024_python.ipynb](#) PYNB

16.5 Ko



[intro_mixte_2024_R.qmd](#) QMD

10.3 Ko



[politeness_data.csv](#) CSV

Il y a 84 observations et 6 variables dans ce tableau :

- `subject` : un identifiant de l'individu
- `gender` : le genre de l'individu
- `scenario` : scénario dans lequel la voix a été enregistré
- `attitude` : le registre formel (`pol`) ou informel (`inf`)
- `frequency` : la fréquence de la voix (en Hz)

1.4 Ko

Un autre jeu de données



[bounce.csv](#) CSV

Il s'agit d'un problème classique de Data Science. On souhaite analyser des données collectées sur d'un site web présentant des recettes de cuisine. le **temps de rebond** est la durée (en secondes) qui sépare le premier accès à une page web du moment où l'utilisateur la quitte.

Pour comprendre cette durée (ici toujours non censurée, elle est courte !), on a réalisé une expérience sur des individus au Royaume-Uni. Deux miracles se produisent ici : (1) c'est l'ordinateur qui mesure la durée, il a une précision à la dixième de milliseconde et (2) il n'y a pas de censure, la durée est relativement courte. On a enregistré :

- `bounce_time` : la durée de rebond de l'individu (en seconde)
- `age` : l'âge de l'individu (en années)
- `county` : le comté dans lequel habite l'individu
- `location` : un code qui dépend du lieu exact d'habitation de l'individu.

Il n'y a que trois valeurs possibles pour `location`, mais 8 comtés différents, donc au total $8 \times 3 = 24$ lieux d'habitation différents.

11.2 Ko



Analyse de survie



[Elements of Survival Analysis](#)

7.8 Mo · Déposé le 21 nov. 24, 15:02



Lab sessions: Survival Analysis



[TP survival: KM and AFT.pdf](#)

28.2 Ko · Déposé le 21 nov. 24, 07:43



[motors.csv](#)

645 octets · Déposé le 21 nov. 24, 07:44



[capacitor.csv](#)

954 octets · Déposé le 21 nov. 24, 07:50



[lung.csv](#)

6.4 Ko · Déposé le 21 nov. 24, 07:46



Évaluation de l'UE

Parcours DS

Projets (groupe de 2 à 4 personnes) :

1. Modèle mixte avec Tensorflow OU
2. Validation croisée avec le score de Brier dans les modèles AFT et Cox OU
3. Résoudre l'exercice pour prédire $T_i - Y_i$ si censure et durée observée = Y_i et covariables X_{ij} OU
4. Mettre en place une démarche de sélection de variables par critère BIC dans les modèles AFT pour le jeu de données Rossi (récidive après libération de prison)

Épreuve écrite : - (voir seconde partie de l'UE avec H. Lorenzo)

Parcours CMB

Projets en binôme (ou trinôme) :

- Choisir un des jeux de données exemple avec plusieurs covariables

- Écrire un rapport d'analyse du jeu de données qui présente : la question, la méthodologie, les données et les conclusions

Epreuve écrite : 1h pendant la semaine du 16 au 20 décembre 2024

Parcours MaSCo

Projets en binome (ou trinôme) :

- Choisir un des jeux de données exemple avec plusieurs covariables
- Écrire un rapport d'analyse du jeu de données qui présente : la méthodologie, les données et les conclusions

Epreuve écrite : 1h pendant la semaine d'examen du 6 au 10 janvier 2025



Retour projet CMB et MaSCo

Ouvert le : jeudi 21 novembre 2024, 00:00 **À rendre** : lundi 13 janvier 2025, 03:00



Non disponible à moins que : Vous ne soyez pas membre de **DS**



[Retour projet DS](#)

Ouvert le : jeudi 21 novembre 2024, 00:00 **À rendre** : lundi 27 janvier 2025, 03:00



Parcours DS, projet 2 : score de Brier

Le score de Brier est utilisé comme critère de qualité d'un modèle AFT ajusté dans des démarches de validation. Voir, par exemple :

https://square.github.io/pysurvival/metrics/brier_score.html

L'objectif de ce projet est de comprendre cette métrique, d'un point de vue théorique, et de montrer son utilisation sur un ou deux exemples, pour faire de la validation ou validation croisée par exemple.



Parcours DS, projet 3 : prédire

En supposant que l'on a ajusté un modèle AFT sur des données, proposer des méthodologies pour prédire :

- la durée T^* d'un nouvel individu dont les covariables sont X^*
- la fonction de survie de T^* pour ce nouvel individu
- la durée T^* d'un nouvel individu dont les covariables sont X et la durée avant censure Y

Tester cette méthodologie sur un ou plusieurs jeu de donnée.



Parcours DS, projet 4 : Données sur la récidive criminelle



[Rossi.csv](#)

137.6 Ko · Déposé le 28 nov. 24, 10:51

Le fichier Rossi.csv contient les données de

Rossi, P.H., R.A. Berk, and K.J. Lenihan (1980). *Money, Work, and Crime: Some Experimental Results*. New York: Academic Press.

Il concerne une étude expérimentale sur l'étude de la récidive criminelle sur 432 hommes des USA à leur sortir de prison. La question d'intérêt principale est de savoir s'il est utile de donner une aide financière à la sortie de prison, pour lutter contre la récidive. Les données contiennent les variables ci-dessous :

- week: durée observée avant la première arrestation après la remise en liberté, ou la censure
- arrest: 1 si la personne a été arrêté à la sortie de prison, 0 sinon
- fin: "yes" si aide financière à la sortie de prison, "no" sinon

- age: âge de l'individu en années, à la date de sortie de prison
- race: "black" ou "other"
- wexp: expérience de travail à temps plein avant l'incarcération qui précède cette étude
- mar: statuts marital à la sortie de prison
- paro: libération conditionnelle
- prio: nombre de condamnation avant celle qui précède cette étude
- educ: niveau d'éducation (2 = 6th grade or less; 3 = 7th to 9th grade; 4 = 10th to 11th grade; 5 = 12th grade; 6 = some college)

Il contient également une variable mesurée chaque semaine dans les colonnes emp1 à emp52. Ainsi emp1 vaut "yes" si la personne a un travail pendant la 1ère semaine avec la remise en liberté, emp1 pendant la 2ème semaine, etc.

Vous devez

- construire une (ou des) covariables numériques pertinentes, non dépendantes du temps, à partir des covariables empX (durée d'emploi, etc.)
- utiliser une méthode progressive ou rétrograde pour ajuster en modèle AFT en sélectionnant les covariables avec un critère BIC
- interpréter les résultats obtenus.



Parcours DS, projet 1 : modèle mixte



[MixedModels Tensorflow.pdf](#)

1.7 Mo · Déposé le 28 nov. 24, 11:26



Lab session: Other datasets for Cox models

Données sur la durée avant l'abandon des études universitaires

- **dur** : durée jusqu'à l'obtention du diplôme ou l'abandon (en mois)
- **evt** : 1 si abandon, 0 sinon
- **sex** : 0 si homme, 1 si femme
- **grd** : note moyenne au lycée (A, la meilleure, à C)
- **prt** : études à temps partiel (1 = oui, 0 = non)
- **lag** : écart en mois entre les études au lycée et à l'université
- **mrg** : date du mariage (en nombre de mois depuis 1/1/1980)
- **stm** : date du début des études (en nombre de mois depuis 1/1/1980)



[Yamaguchi.csv](#)

5.6 Ko · Déposé le 5 déc. 24, 12:14

Données biographiques allemandes

- **id** : identifiant de l'individu
- **sn** : numéro de l'événement (emploi)
- **ts** : date de début
- **tf** : date de fin
- **sex** : genre de l'individu
- **ti** : date de l'interview
- **tb** : date de naissance
- **te** : date d'entrée sur le marché de l'emploi
- **tmar** : date du mariage (0 si non marié)
- **pres** : prestige de l'emploi
- **presn** : prestige de l'emploi suivant

- edu : plus haut niveau d'éducation atteint
- des : état de destination
- ttfp : temp d'emploi
- lfx : expérience sur le marché de l'emploi
- pnoj : nombre d'emplois précédents
- cohort : 1929-1931 ou 1939-1941 ou 1949-1951
- coho1, 2, 3 : indicatrices des trois cohortes

Les dates sont en mois depuis le 1/1/1900



[biographie_allemande.csv](#) 

43.0 Ko · Déposé le 5 déc. 24, 12:39